Know-Center at PAN 2015 author identification
Notebook for PAN at CLEF 2015

Oliver Pimas, Mark Kröll, and Roman Kern

Know-Center GmbH
Graz, Austria
{opimas, mkroell, rkern}@know-center.at

Abstract Our system for the PAN 2015 authorship verification challenge is based upon a two step pre-processing pipeline. In the first step we extract different features that observe stylometric properties, grammatical characteristics and pure statistical features. In the second step of our pre-processing we merge all those features into a single meta feature space. We train an SVM classifier on the generated meta features to verify the authorship of an unseen text document. We report the results from the final evaluation as well as on the training datasets.

1 Introduction

The paper at hand presents a detailed description of our approach to solve the author identification task at PAN 2015. The problem to solve can be formulated as follows: Given a set of documents by a single known author as well as a document of unknown authorship, determine whether this unknown document was written by that particular author or not. This problem is also labelled authorship verification. For the PAN 2015, the training set for a single author consisted of text documents from different genres and different topics. Therefore, the task can be seen as cross-genre and cross-topic authorship verification. This resembles real-world applications more closely but also makes the task more challenging.

This notebook paper is outlined as follows: In section 2 we describe our classification approach. In section 3 we present the results. These are followed by a conclusion in section 4.

2 Approach

We based our work for the PAN 2015 author identification challenge upon Know-Center submissions of previous years (see [5], [6], [7]). We consider authorship verification a supervised classification problem. For each author, we pre-process each document in two steps. In step one we extract different features. These features include statistical features such as term frequencies, character or word n-grams. We also extract grammar features such as possible wrong quotes, unpaired brackets or sentences starting with upper-case letters. Stylometric features including Hapax Legomena [11], Brunets W [11], Simpsons D [11], Sichels S [11] or Honores H [11] or sentence length n-grams try to capture the writing style of an author. We also extract topic features in order to

model the topics an author tends to write about. However, as the PAN 2015 challenge was cross-topic, we deactivated the topic features for our final evaluations.

For most of the used features please refer to the original paper [7] as well as the follow-up submissions [6] and [5]. One of the new features is sentence length n-grams. We define sentences consisting of up to 7 words as short sentences. Sentences consisting of more than 13 words are considered long sentences. We consider sentences that neither qualify as short or long sentence to be medium sentences. Using this definitions, we move a window of size n over an author's text and store n-grams as features, substituting the sentence with a length indication character. A short sentence is substituted by *s*, a medium sentence by *m* and a long sentence by *l*. Thus, a sentence length bigram describing a short sentence followed by a long sentence would simply be "*sl*". We store sentence length n-gram as frequency vectors. The intuition behind this feature is to model an authors tendency to write longer or shorter sentences or to mix sentence length.

Another new feature is topic distribution. We extract the topic distribution of the training corpus using MALLET's [8] implementation of LDA [1]. LDA is a generative model that tries to uncover latent topics from a given text corpus. MALLET's *ParallelTopicModel*[1] is a parallel threaded implementation of LDA building upon the work of [9] and [12]. We generate a topic model for each language, if multiple languages are present in the training set. In the feature extraction step, we then store the topic distribution vector of every document for each author. However, since the authorship verification challenge at PAN 2015 was cross-topic, we deactivated this feature set for the final evaluation.

The grammar features are extracted using the open source style and grammar checker LanguageTool[2]. Please refer to [5] for more details.

The PAN 2015 authorship verification challenge includes datasets from the languages English, Spanish, Dutch and Greek. While our preprocessing pipeline generally supports all four languages, not all features can be extracted for each languages. We currently do not support stemming, stop or function word annotation or part-of-speech annotation for Dutch and Greek. Therefore, we cannot extract all features for those two languages and results may differ.

After extracting those features in step one, we face a number of different feature spaces with different ranges of values. This introduces a problem for many machine learning algorithms. In step two of our pre-processing pipeline, we tackle this problem by generating meta features from those extracted features. These meta features do all exist in a single meta feature space. To generate the meta features, we aggregate the extracted feature spaces and compare it to the unseen document using the Kolmogorov-Smirnov test. For more details on the comparison and the meta feature generation process, please refer to [5].

Finally, we train a classifier on our meta features. For the evaluation we trained an SVM, using the machine learning framework WEKA [2]. We used WEKA's class *SMO*[3], which is an implementation of John Platt's sequential minimal optimization al-

[1] http://mallet.cs.umass.edu/api/cc/mallet/topics/ParallelTopicModel.html

[2] https://languagetool.org/

[3] http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html

gorithm [10] for training support vector classifier, building upon the work of [4] and [3]. We did not evaluate different settings for the support vector classifier but used the default parameter settings of WEKA instead.

3 Results

We report the results on the training and test datasets provided by the PAN 2015 authorship verification challenge. In order to evaluate the performance on the training dataset we used the method *crossValidateModel* provided by the WEKA class *Evaluation* to report the results doing 10-fold cross validation. The results on the final training sets can be seen in table 1. Evaluations on previously released training datasets scored similar results. Especially the performance on the English dataset, where all feature-sets are supported by our pre-processing pipeline (see section 2), look quite promising. The results of the PAN 2015 authorship verification evaluations can be seen in table 2. Comparing the final evaluation results to those on the training datasets it seems our approach is prone to overfit the training dataset.

Dataset	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**AUC**
English training dataset	0.91	0.087	0.912	0.91	0.91	0.882
Spanish training dataset	0.74	0.287	0.743	0.74	0.741	0.595
Dutch training dataset	0.72	0.295	0.72	0.72	0.72	0.634
Greek training dataset	0.74	0.263	0.74	0.74	0.74	0.669

Table 1. The weighted average results on the training datasets as printed by WEKA's *crossValidateModel* method provided by the class *Evaluation*[5].

Dataset	**AUC**	**C1**	**Final Score**	**Runtime**
English testing dataset	0.50692	0.506	0.2565	00:07:21
Spanish testing dataset	0.49	0.49	0.2401	00:04:12
Dutch testing dataset	0.50815	0.51515	0.26178	00:02:27
Greek testing dataset	0.48	0.48	0.2304	00:03:57

Table 2. The results as reported from the evaluation runs of the PAN 2015 authorship verification challenge. Separate models were trained on the training datasets for each language. Compared to our training results, it seems our models overfitted the training data.

4 Conclusion

We presented our system developed for the PAN 2015 authorship verification challenge. Our system is based on several different feature spaces that are combined into a single

[5] http://weka.sourceforge.net/doc.dev/weka/classifiers/Evaluation.html

meta feature space in a two-step preprocessing pipeline. Our systems performance in the final evaluation did not meet our expectations. Comparing the results on the training and test datasets, our system seems to overfit the training data.

4.1 Future Work

In the future we aim to invest into trying different combinations of features as well as tuning the classification model creation. We also plan to try different supervised classification algorithms and compare the results. In order to validate our meta feature generation approach, we plan to use a classification algorithm that is able to deal with different feature spaces and compare the results of this algorithm, using the features extracted in step one of our pre-processing pipeline, to those of our classifier trained on the meta features generated on pre-processing step two.

5 Acknowledgements

This work is funded by the KIRAS program of the Austrian Research Promotion Agency (FFG) (project number 840824). The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. The Journal of Machine Learning (2003)
2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software : An Update. SIGKDD Explorations 11(1), 10–18 (2009)
3. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems. vol. 10. MIT Press (1998)
4. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to platt's smo algorithm for svm classifier design. Neural Computation 13(3), 637–649 (2001)
5. Kern, R.: Grammar Checker Features for Author Identification and Author Profiling. CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers (2013), http://ims-sites.dei.unipd.it/documents/71612/430938/CLEF2013wn-PAN-Kern2013.pdf
6. Kern, R., Klampfl, S., Zechner, M.: Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification - Notebook of PAN at CLEF 2012. Working Notes Papers of the CLEF 2012 Evaluation Labs pp. 1–15 (2012)
7. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/Veto Meta-Classifier for Authorship Identification - Notebook for PAN at CLEF 2011. CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands (2011)
8. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu

9. Newman, D., Asuncion, A., Smyth, P., Welling, M.: Distributed Algorithms for Topic Models. The Journal of Machine Learning Research 10, 1801–1828 (2009), http://dl.acm.org/citation.cfm?id=1577069.1755845
10. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998), http://research.microsoft.com/~jplatt/smo.html
11. Tweedie, F.J., Baayen, R.H.: How Variable May a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities 32(5), 323–352 (1998), http://link.springer.com/article/10.1023/A%3A1001749303137
12. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09 p. 937 (2009), http://portal.acm.org/citation.cfm?doid=1557019.1557121